How do Users Handle Inconsistent Information? The Effect of Search Expertise

Kazutoshi Umemoto*, Takehiro Yamamoto, and Katsumi Tanaka
Kyoto University, Yoshida Honmachi, Sakyo, Kyoto, 606-8501, Japan
{umemoto,tyamamot,tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

While search engines sometimes return different documents containing contradictory answers, little is known about how users handle inconsistent information. This paper investigates the effect of *search expertise* (defined as specialized knowledge on the internal workings of search engines) on search behavior and satisfaction criteria of users. We selected four tasks comprising factoid questions with inconsistent answers, extracted answers that 30 study participants had found in these tasks, and analyzed their answer-finding behavior in terms of the presence or absence of search expertise. Our main findings are as follows: (1) finding inconsistent answers causes users with search expertise (search experts) to feel dissatisfied, while effort in searching for answers is the dominant factor in task satisfaction for those without search expertise (search non-experts); (2) search experts tend to spend longer completing tasks than search non-experts even after finding possible answers; and (3) search experts narrow down the scope of searches to promising answers as time passes as opposed to search non-experts, who search for any answers even in the closing stage of task sessions. These findings suggest that search non-experts tend to be less concerned about the consistency in their found answers, on the basis of which we discuss the design implications for making search non-experts aware of the existence of inconsistent answers and helping them to search for supporting evidence for answers.

CCS Concepts

•**Information systems** → **Users and interactive retrieval;** *Question answering;* •**Human-centered computing** → *User studies;*

Keywords

search expertise; satisfaction; factoid questions; answer consistency

*Research Fellow of Japan Society for the Promotion of Science

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

SAC 2016, April 04 - 08, 2016, Pisa, Italy

Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 978-1-4503-3739-7/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2851613.2851698

1. INTRODUCTION

People sometimes conduct complex search tasks [24] where their information needs are not satisfied by a single search query. Multiple queries can be required to complete a task even if the task comprises a single search goal [13]. Take, as one such example, a user who queries a search engine about the worst year for droughts in US history, expecting a single answer. If the engine returned different search results giving different years as the answer to the query, this user would be unsure about what the correct answer was and might finish the task dissatisfied. To fulfill the goal of this task, more queries would need to be issued and more documents assessed so that he/she could collect possible years with supporting evidence and judge the most likely one as the task answer.

The issue in the above example is not limited to *relevance* [3], which has been studied for decades, but more complex notions, like *credibility* [6], might also play important roles in such intellectual activities. As explained in the literature [26], the credibility of information is a subjective quality whose interpretation depends on its receivers. If users do not understand the underlying algorithms of search engines, they might trust answers described in top-ranked documents as credible and finish their task without in-depth exploration. While specialized knowledge of users has been shown to affect their search strategies and outcome evaluations [11, 22, 26], little is still known about how answer inconsistency affects them.

We investigated how users handle inconsistent information found in search tasks, focusing on *search expertise*, which we defined as specialized knowledge on the internal workings of search engines, as a possible factor affecting user behavior and satisfaction. The research questions addressed in this paper are: (1) how does the inconsistency of answers that users find relate to the degree of task satisfaction that they gain? (2) when do users complete search tasks where inconsistent answers are found? and (3) how do answers targeted by users change as time passes? To answer these questions, we selected four tasks comprising factoid questions [4] where answers obtained through search engines are inconsistent with each other, extracted answers that 30 study participants had found in these tasks, and analyzed answer-finding behavior of the participants in terms of the presence or absence of search expertise.

The main findings from our analysis can be summarized as follows: (1) finding inconsistent answers caused dissatisfaction for users with search expertise (search experts), while effort in searching for answers was the dominant factor in task satisfaction for those without search expertise (search non-experts); (2) search experts tended to spend longer completing tasks than search non-experts even after finding possible answers; and (3) search experts narrowed down

the targeted information to promising answers as time passed as opposed to search non-experts, whose found answers were still diverse even in the closing stage of their task sessions. On the basis of these findings, which suggest search non-experts care little about the consistency in their found answers, we discuss the design implications for making search non-experts aware of the existence of inconsistent answers and helping them to search for supporting evidence for answers.

2. RELATED WORK

Existing studies related to our work fall into the following research areas: (1) question answering (QA) on the Web, (2) evaluating search outcomes from the perspective of real users, and (3) understanding user-dependent factors that affect search behavior, each of which is overviewed in turn below.

Question Answering. People often use the Web to find answers to their questions [16, 17]. To foster research on systems that directly return correct answers for a given question, the QA track was held in Text REtrieval Conferences (TRECs)[1] from 1999 to 2007. The most dominant questions used in the TREC 2007 QA track were categorized into the *factoid* type, which asked for a fact-based short answer. As discussed by Dang *et al*. [4], different documents might support contradictory answers as being correct. While the main focus of the QA track was on QA systems, less attention has been paid to real users, particularly on how they handle such inconsistent answers during their searches.

Search Satisfaction. Batch-style evaluation, which is based on relevance judged by assessors, has been reported to produce different results from user-oriented evaluation, in which the effectiveness of a system is measured through its use by real users [19, 27]. In light of this mismatch, researchers have recently directed their attention to evaluating search outcomes in more direct ways. Particularly, predicting user satisfaction from search behavior is becoming a trending research topic [1, 9, 14, 20], while the granularity of satisfaction to be predicted differs from work to work (*e.g.*, task-level [1] and click-level [14]). Wang *et al*. [20] modeled latent satisfaction for individual actions to predict task satisfaction. Their model outperformed state-of-the-art satisfaction prediction models, suggesting the importance of considering fine-grained actions to predict task satisfaction. Effort expended to complete tasks has shown to be related to user satisfaction [7, 8, 25]. Guo *et al*. [8] reported a negative correlation between task satisfaction and effort features (*e.g.*, the numbers of issued queries and browsed pages). Time spent completing tasks has also been shown by Xu and Mease [25] to be negatively correlated with user satisfaction.

User Factors. Much effort has been devoted to analyzing search behavior logs in order to better understand how users engage in search activities [2, 10, 11, 21, 22]. Some analytical studies have pointed out the effect of user-dependent factors on their search behavior. Hölscher and Strube [11] analyzed the difference in query formulation strategies between users with high search skills and other users. Their analysis showed that the former users used advanced search options more frequently than the latter. Knowledge of search domains, or domain expertise, is known to influence the likelihood of search success [22]. White [21] investigated how the beliefs and biases of users affected their search behavior. His surveys, targeting yes-no questions in the medical domain, revealed that users had the search biases of preferring information affirming their beliefs, which led them to settle on incorrect answers about half the time.

Building on existing research, the present work studies how users handle inconsistent information. Search tasks comprising factoid questions were targeted for our analysis. We investigated answers found by users as well as search effort expended to complete tasks as possible factors affecting task satisfaction. We analyzed the difference between users with and without search expertise, in terms of their answer-finding behavior and satisfaction criteria.

3. DATASET

We adopted as a source of our analysis the publicly available search logs[2] of a user study [5]. Four tasks were selected from this source, in which inconsistent information existed as the task answers. We manually checked Web pages that study participants had browsed in these tasks to identify their found answers.

3.1 User Study

Feild *et al.* [5] conducted their study in the fall of 2009, aiming at predicting frustration of users from their search behavior. Given the purpose of frustration prediction, all the 12 search tasks in this study were designed to be difficult to achieve (*i.e.*, they could not be completed just by browsing a single Web page). As study participants, 30 people (23 males and 7 females) were recruited from diverse departments (computer science, English, kinesiology, physics, *etc*.) of a university. Each participant was asked to conduct seven search tasks one by one, whose ordering was determined by the Latin square design so as to remove task ordering effects. The participant's activities (including issued search queries and browsed Web pages) were logged with the timestamps during each task session. A post-questionnaire was administered after each task was executed, in which participants reported the following items on the basis of their search outcomes: (1) the degree to which their information needs were satisfied (on a five-point scale: $1 =$ not satisfied at all; $5 =$ completely satisfied) and (2) answers that they found and judged as likely. Further information on this user study can be found elsewhere [5].

3.2 Task Selection

Three search tasks of the above-mentioned user study (*e.g.*, "Name three bridges that collapsed in the USA since 2007.") could be categorized into *list* questions [4], which asked for distinct, multiple instances satisfying the information need. These were excluded from our analysis as users would anticipate the existence of multiple answers in such tasks. This left nine tasks corresponding to *factoid* questions [4], which asked for a single answer. These tasks were relevant to our interests as different documents might support contradictory answers as being correct. We decided, however, to exclude four of them comprising question templates (*e.g.*, "Find the hours of the PetsMart nearest [Wichita, KS; Thorndale, TX; Nitro, WV]."), because different variants for an identical task had different answers, which would make our analysis more complicated. One task whose correct answer was time-dependent (*i.e.*, "How much did the Dow Jones Industrial Average increase/decrease at the end of yesterday?") was also excluded for similar reasons. In this way, the remaining four tasks were targeted for our analysis: *Drought*, *Pixels*, *TV*, and *Verizon*. Table 1 summarizes the descrip-

[1] http://trec.nist.gov/data/qamain.html

[2] http://hank.feild.net/downloads.html

Table 1: Task descriptions given to participants of user study.

Task	Description
Drought	In what year did the USA experience its worst drought? What was the average precipitation in the country that year?
Pixels	How many pixels must be dead on a MacBook before Apple will replace the laptop? Assume the laptop is still under warranty.
TV	What was the best selling television (brand & model) of 2008?
Verizon	What's the helpline phone number for Verizon Wireless in MA?

Table 2: Statistics on task answers found by study participants with answer samples.

	Drought	Pixels	TV	Verizon
Answer samples	1930 to 1931	Any pixels	Samsung	800-922-0204
	1950s	5 pixels	LN32B460	800-899-4249
	1988 to 1989	Case by case	LN52A650	800-256-4646
	2001 to 2003	No public policy	Sony	1-800-VERIZON
# of answers	33	17	5	7
Cohen's κ	0.75	0.52	0.78	1.00

tions of these tasks given to study participants. As shown at the top of Table 2, participants did find inconsistent answers in these tasks.

3.3 Answer Extraction

While our interests lie in the effect of inconsistent answers on answer-finding behavior of users, the original logs contained only the answers that participants believed to be most likely. To obtain an entire set of answers that they found through their sessions of the four targeted tasks, two of the authors independently extracted answer entities (simply referred to as answers hereinafter) from each page that they browsed. Note that some pages could have more than one answer. We excluded search engine results pages (SERPs) from extraction sources for two reasons: (1) users do not always scan all titles and snippets in SERPs; and (2) these descriptions are sometimes insufficient for answer evidence due to being short and incomplete. To make the extracted answers as consistent as possible, the assessors shared a clear criterion for answer extraction; descriptions in non-SERPs should be extracted if and only if they, at least partly, answer the task questions listed in Table 1. Table 2 summarizes the inter-assessor agreement scores of extracted answers in terms of Cohen's κ. When calculating these scores, we regarded pages from which both (or neither) of the assessors extracted answers as accordantly judged. The score across all four tasks was 0.78, which could be regarded as substantial agreement according to the literature [15]. A similar tendency was observed for each task except for the Pixels task, where the score corresponded to moderate agreement ($\kappa = 0.52$). As can be seen from the samples of extracted answers in Table 2, the Pixels task did not have common notation patterns as its answers unlike other tasks, which could be a possible reason for its comparatively low score. In addition, the assessors might miss some answers to be extracted as they were not native speakers of English, which was the language used in the target pages. To deal with this problem, we decided to use in our analysis a union of answers extracted by the assessors as the ground truth of answers that participants found.

4. ANALYSIS

This section describes our analysis on how searchers handle inconsistent information. In light of the argument described in Section 1, we focused on search expertise as a possible factor affecting their search behavior and outcomes, where search expertise was defined as specialized knowledge on the internal workings of search engines. Given the information on participants' major fields logged in the original dataset, we regarded those who studied information retrieval as having search expertise. Hereinafter, we will refer to users with search expertise as *search experts* and those without it as *search non-experts*[3]. Table 3 shows the assignments of our targeted tasks to study participants. There were 66 task sessions in total, of which 17 ($\simeq 26\%$) were done by 7 search experts and 49 ($\simeq 74\%$) by 23 non-experts. Participants performed at least two tasks on average.

When testing significant effects in our analysis, we used the significance level of $\alpha = .05$. On ground of the small sample size of our dataset, however, the phrase "a trend toward significance" is also used when the p-value of a significance test falls within the range from .05 to .10. Overall, the means of task satisfaction reported by search experts and non-experts were 3.65 and 3.47, respectively. The difference was shown to be insignificant by Welch's t-test ($t(41) = .704,\ p = .485$), which suggests that both experts and non-experts felt, on average, almost the same degree of satisfaction. To explore in detail the effect of search expertise on answer-finding behavior, we analyzed the relationship between found answers and task satisfaction (Section 4.1), the distribution of time spent finding answers and completing tasks (Section 4.2), and the change in participants' targeted answers over time (Section 4.3).

4.1 How do Answers Relate to Satisfaction?

We began by analyzing the effect of search expertise on task satisfaction. As effort expended to complete tasks has shown to be negatively correlated with search experience [2, 8], we formulated the following hypothesis on the relationship between search effort and task satisfaction.

Hypothesis 1 (H1): *Increased search effort leads to dissatisfaction with the task regardless of the presence or absence of search expertise.*

Given the characteristics of our targeted tasks, where inconsistent information was found from different documents, users might feel dissatisfied when finding different answers. We hypothesized that search expertise played an important role in the relationship between answer consistency and task satisfaction. Our second hypothesis is summarized as follows:

Hypothesis 2 (H2): *Only search experts show dissatisfaction with the task when finding many inconsistent answers.*

4.1.1 Procedure

We considered two effort features for the hypothesis **H1**: (1) the number of issued queries and (2) the number of browsed pages. As for the hypothesis **H2**, we designed three features to measure the answer consistency: (1) the total number of found answers, (2) the number of unique found answers, and (3) the entropy of found answers. These three features are calculated as follows: (1) $\sum_{a \in A} m(a)$, (2) $|A|$, and (3) $-\sum_{a \in A} \frac{m(a)}{\sum_{a' \in A} m(a')} \log_2 \frac{m(a)}{\sum_{a' \in A} m(a')}$, where A is a set of answers found by a user in a session of a certain task, and $m(a)$ denotes how many times an answer $a \in A$ was found by the user in the session. We calculated these features for each session in our dataset and used Pearson's r between each feature and task satisfaction to test the two hypotheses.

[3]The term "search" will be omitted if it is clear from the context.

Table 3: Task assignments to participants.

	Experts							Non-Experts																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Drought		✓	✓	✓		✓	✓	✓		✓	✓	✓			✓	✓				✓		✓			✓		✓			✓
Pixels	✓		✓	✓	✓	✓				✓		✓		✓			✓			✓	✓		✓			✓				✓
TV		✓	✓		✓	✓		✓	✓		✓	✓	✓	✓	✓			✓	✓	✓	✓	✓					✓	✓	✓	
Verizon		✓		✓			✓	✓	✓	✓			✓	✓	✓		✓	✓	✓			✓		✓	✓	✓		✓		

Table 4: Features related to search effort and found answers.

	Experts		Non-Experts	
Feature	mean (SD)	r (p-value)	mean (SD)	r (p-value)
# of queries	3.06 (1.39)	−.266 (.303)	2.53 (1.23)	−.546 (.000)
# of pages	4.88 (2.23)	−.631 (.007)	4.02 (2.38)	−.072 (.625)
# of total answers	4.18 (3.13)	−.380 (.132)	3.53 (2.34)	.114 (.435)
# of unique answers	3.71 (2.76)	−.339 (.183)	3.02 (1.84)	.181 (.213)
Answer entropy	1.69 (0.85)	−.458 (.074)	1.59 (0.73)	−.021 (.896)

4.1.2 Results

Table 4 lists the statistics of these features for task sessions grouped by search expertise. The value on the left of each cell in the table is the mean (with standard deviation) for the feature values, and that on the right is Pearson's r (with p-value) between the feature and satisfaction. We will first describe the feature trend common in both user groups. We will then discuss the effect of search expertise on task satisfaction.

A general trend we found from Table 4 was that the effort features tended to be negatively correlated with task satisfaction irrespective of search expertise, which confirms our hypothesis **H1**. This result is consistent with the previous work that showed the negative effect of user effort on search performance [2, 8]. However, Table 4 also indicates another finding that the most dominant effort feature depends on the presence or absence of search expertise; the number of pages significantly correlated with task satisfaction only for search experts ($r = -.631,\ p = .007$), while the number of queries was significant only for non-experts ($r = -.546,\ p < .001$). Search experts are, by our definition, familiar with the internal functions of search engines, such as query processing and document ranking. We conjecture that having such technical knowledge helps them to formulate effective search queries with less effort than others (as discussed by Hölscher and Strube [11]), which may reduce the load of querying and result in the low effect of the query effort feature.

In contrast to the effort features, which tended to be similar among the participants, we can observe from Table 4 that the answer features seem to affect task satisfaction differently depending on search expertise. The correlation between these features and task satisfaction tends to be negative for search experts. In particular, answer entropy showed a trend toward significance for them ($r = -.458,\ p = .074$). For non-experts, however, none of these features significantly correlates with task satisfaction. This result supports our hypothesis **H2** on the difference in satisfaction criteria between search experts and non-experts. A possible reason for the difference is their degree of trust in search engines. According to an online survey conducted by Nakamura *et al*. [18], many general users trust the ranking of search results to some extent despite insufficiently understanding the underlying mechanism of Web search engines, which suggests that search non-experts may have an over-reliance on top documents in the ranking. We infer that specialized knowledge on search engines prevents search experts from relying excessively on top documents and encourages them to examine the consistency of obtained answers by expending additional effort.

4.2 When Do Users Finish Answer Searches?

The analysis in Section 4.1 suggested that search experts might expend more effort than non-experts to examine the answer consistency. Aiming at better understanding on how users expend effort to complete tasks, we formulated and tested the following hypothesis on time spent completing tasks:

Hypothesis 3 (H3): *Search experts take longer to complete tasks than non-experts.*

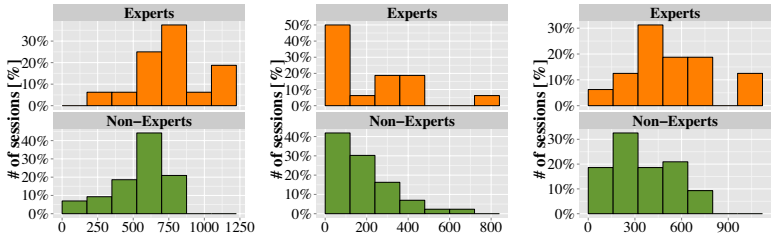
4.2.1 Procedure

To test the hypothesis above, we first analyzed how *completion time*, which is time spent completing individual tasks, differed between search experts and non-experts. In addition to completion time, the following two time-related measures were taken into account in our analysis to reveal in which phases users spent much time during their sessions: *discovery time*, which is time spent finding the first answer; and *remaining time*, which is time elapsing from the first answer discovery to the task completion. Note that remaining time can be obtained by subtracting discovery time from completion time. Completion time has been reported to be negatively correlated with task satisfaction [25]. Shorter discovery time may indicate better search skills, while remaining time may reflect the degree to which users care about their found answers. We calculated these three measures for each session and compared the difference in the time distributions of each measure between search experts and non-experts. Welch's t-test was applied for testing whether two distributions were significantly different.

4.2.2 Results

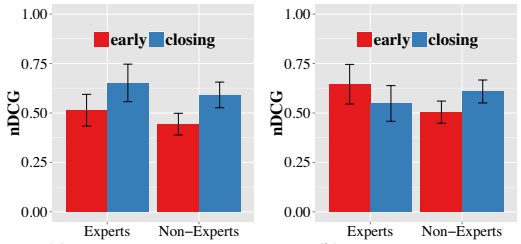
Figures 1(a), 1(b), and 1(c) show the histograms of task sessions with respect to completion time, discovery time, and remaining time, respectively. The top and bottom histograms in each figure represents the session distribution for search experts and search non-experts, respectively.

We can find from Figure 1(a) that search experts tended to spend longer completing the tasks than non-experts. More specifically, about half of the sessions by search experts lasted for more than 10 minutes, while less than one quarter of the sessions by non-experts lasted as long. Welch's t-test revealed that the completion time was significantly longer for search experts than non-experts ($t(24) = 3.29,\ p = .003$), which supported our hypothesis **H3**. We plot the histograms of discovery time in Figure 1(b) to better understand the reason for the search experts' long sessions. While search experts, on average, found the first answers faster than non-experts, Welch's t-test did not reveal a significant difference in the time distributions between these two groups ($t(20) = -.985,\ p = .337$). As shown in Figure 1(c), our analysis on remaining time revealed that search experts tended to spend longer completing tasks than non-experts after finding the first answers, which was shown to be significant



(a) **Completion time [sec]** (b) **Discovery time [sec]** (c) **Remaining time [sec]**

Figure 1: Histograms for number of sessions w.r.t time spent finding answers.



(a) **Submitted answers** (b) **Ignored answers**

Figure 2: Changes in targeted answers over time.

by Welch's t-test ($t(21) = 2.24,\ p = .036$).

Our main finding from the analysis above is that search experts tended to spend longer completing tasks than non-experts primarily due to the relative increase of remaining time. Cautious attitudes by search experts can be inferred from this result. That is, they did not seem to finish search tasks just by finding possible answers. Taking into account the expert-specific characteristic reported in Section 4.1 (*i.e.*, the negative correlation between answer entropy and task satisfaction), we conjectured that search experts spared more time and effort to find evidence that supported the answers they had found.

4.3 How do Users Change Targeted Answers?

The analysis results in Section 4.2 suggested the possibility that search experts spent a long time finding evidence that supported their found answers. To confirm whether this conjecture was correct, we investigated the change in answers targeted by users over time. The following hypothesis was tested in this investigation:

Hypothesis 4 (H4): *Search experts narrow down the search scope of answers as time passes.*

4.3.1 Procedure

To grasp the temporal trends in found answers, we categorized them into the following two classes: *submitted answers*, which contained answers that users found and submitted as their responses for the post-task questionnaire; and *ignored answers*, which contained answers they found but did not submit. The ignored answers for a session were extracted by eliminating the submitted answers from all the answers found in the session. For each of these classes, we measured how the corresponding answers were found in the *early* and *closing* stages of sessions in the following manner. First, we treated a sequence of browsed pages, except SERPs, in each session in chronological order as the early stage and those in reverse chronological order as the closing stage[4]. Then, we calculated normalized Discounted Cumulated Gain (nDCG) [12] for the sequence of each stage by regarding pages with answers belonging to the class as relevant. Note that a session receives a high nDCG score when answers of the targeted class are frequently found in the targeted stage. We applied paired t-test to test whether nDCG scores significantly differed between the early and closing stages.

4.3.2 Results

We show the mean nDCG scores for the early and closing stages in Figure 2. The results for submitted answers are shown in Figure 2(a) while those for ignored answers in Figure 2(b). The error bar on each bar chart represents the standard error of the mean.

As shown in Figure 2(a), we observed a common trend between search experts and non-experts for submitted answers; the mean nDCG score tended to increase as time passed for both user groups (from .514 to .652 for search experts and from .444 to .591 for non-experts). These changes were shown by paired t-test to be an increasing trend toward significance for search experts ($t(16) = 1.75,\ p = .099$) and to be a significant increase for non-experts ($t(48) = 3.50,\ p = .001$). As for ignored answers, on the other hand, we observed different trends between search experts and non-experts as shown in Figure 2(b); the mean nDCG score tended to decrease from .646 to .548 for search experts while tended to increase from .504 to .609 for non-experts. Applying paired t-test revealed the decrease for search experts as insignificant ($t(16) = -1.26,\ p = .225$) and the increase for non-experts as significant ($t(48) = 2.07,\ p = .044$). These results indicated that only search experts tended to search for documents with submitted answers more intensively than those with ignored answers at the closing stage of their sessions, which supports our hypothesis **H4**.

We inferred from these findings that search experts may attempt to find other sources that supported answers that they had found to date by narrowing down the scope of searches to these answers. This behavior could also be interpreted as a sign of their cautious attitudes to found answers, as was described in Section 4.2. In contrast, our results also suggested that search non-experts seemed to search for any kind of answer, irrespective of its class, even in the closing stage of their sessions. The difference between search experts and non-experts may imply that search expertise affected whether and how users validated their found answers, which is discussed in more detail in the remainder of this paper.

5. DISCUSSION

We have shown that the absence of search expertise leads users (1) to judge task satisfaction mainly by search effort that they expend, (2) to finish search tasks without in-depth exploration, and (3) not to narrow down the targeted information to promising answers even in the closing stage of their task sessions.

5.1 Implications

Our findings suggest that search non-experts tended to be less concerned about the consistency in their found answers. Possible explanations for this trend are: (1) they were unaware of the existence of inconsistent answers; (2) they were unwilling to conduct in-depth exploration as additional effort would have been required; and (3) they intended to conduct in-depth exploration but failed due to their lack of knowledge or skills. To help search non-experts facing the situations above to properly handle inconsistent information, search engines first need to notify them of the existence of

[4]We also tried dividing browsed pages into the two stages with some empirically defined cutoffs and obtained similar results, which are omitted from this paper due to space limitations.

multiple instances as possible answers. However, just presenting all the instances may cause them to feel apprehensive about the real answer, similar to the way that search experts felt dissatisfied when finding diverse answers (Table 4). Such apprehension can be reduced by providing complementary information (*e.g.*, majority and typicality) for each answer. Ranking documents containing these answers is also important as many users have been shown to interpret search results as an ordering of likelihood [23]. In addition, search engines need to provide functions for in-depth exploration of each answer. It would be worth establishing techniques to retrieve evidence supporting a given answer and measure its reliability, which would help search non-experts to find the most credible answer with less apprehension.

5.2 Limitations

There are, at least, four limitations of our methodology that we should acknowledge. First, we had to use a small dataset comprising 66 task sessions for our analysis, given the effort required for answer extraction. As this dataset was sampled from search logs in a user study [5] designed for a different purpose, search experts and non-experts were not spread evenly across tasks as was shown in Table 3. A purpose-built study with large, balanced samples would be needed to confirm our findings and gain more insight into answer-finding behavior. Second, the present work focused only on factoid questions to simplify our analysis. There are, however, other types of questions (*e.g.*, list and definition [4]). It is also worth investigating the effect of search expertise on answer-finding behavior for such question types and identifying the differences from those for the factoid type. Third, two assessors were not native speakers of English, which was the language of Web pages from which they extracted answers that study participants had found. Although we used, given the inter-assessor agreement, a union of their extracted answers as the ground truth to avoid answers being missed, this issue may have affected our analysis. The last limitation is about our definition of search expertise. The present work defined search expertise as specialized knowledge on the internal workings of search engines. This definition may be too strict because users who have been using search engines for years are likely to have experienced search skills even if they do not have the specialized knowledge. Furthermore, while we analyzed the effect of search expertise on search behavior and outcomes, our analysis results might also be affected by domain expertise [22].

6. CONCLUSIONS

This paper investigated the effect of search expertise on search behavior and satisfaction criteria of users. Targeting four tasks comprising factoid questions with inconsistent answers, we analyzed answer-finding behavior of 30 study participants in terms of the presence or absence of search expertise. Our findings suggested that search non-experts tended to be less concerned about the consistency in their found answers. To help search non-experts to properly handle inconsistent information, search engines need to notify them of the existence of multiple answer instances and establish techniques for retrieving evidence that supports a given answer.

Future work includes further investigating the effect of search expertise by conducting a purpose-built study with large, balanced samples. While we focused only on factoid questions in the present work, we are also interested in answer-finding behavior for other question types (*e.g.*, list and definition). It would also be worth exploring not only search expertise but also other user factors, like domain expertise, as possible factors affecting answer-finding behavior and search outcomes of users.

7. ACKNOWLEDGMENTS

This work was supported in part by JSPS Grants-in-Aid for Scientific Research (Nos. 15H01718 and 13J06404).

8. REFERENCES

[1] M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: A game for modeling different types of web search success using interaction data. In SIGIR, pp. 345–354 (2011).

[2] A. Aula, R. M. Khan, and Z. Guan. How does search behavior change as search becomes more difficult? In CHI, pp. 35–44 (2010).

[3] S. Büttcher, C. L. A. Clarke, and G. V. Cormack. *Information retrieval: implementing and evaluating search engines*. MIT Press (2010).

[4] H. T. Dang, D. Kelly, and J. Lin. Overview of the TREC 2007 question answering track. In TREC (2007).

[5] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In SIGIR, pp. 34–41 (2010).

[6] B. J. Fogg and H. Tseng. The elements of computer credibility. In CHI, pp. 80–87 (1999).

[7] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2), pp. 147–168 (2005).

[8] Q. Guo, D. Lagun, and E. Agichtein. Predicting web search success with fine-grained interaction data. In CIKM, pp. 2050–2054 (2012).

[9] A. Hassan and R. W. White. Personalized models of search satisfaction. In CIKM, pp. 2009–2018 (2013).

[10] A. Hassan, R. W. White, S. T. Dumais, and Y.-M. Wang. Struggling or exploring?: Disambiguating long search sessions. In WSDM, pp. 53–62 (2014).

[11] C. Hölscher and G. Strube. Web search behavior of internet experts and newbies. *Computer Networks*, 33(1-6), pp. 337–346 (2000).

[12] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4), pp. 422–446 (2002).

[13] R. Jones and K. L. Klinkner. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In CIKM, pp. 699–708 (2008).

[14] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In WSDM, pp. 193–202 (2014).

[15] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), pp. 159–174 (1977).

[16] Q. Liu, E. Agichtein, G. Dror, Y. Maarek, and I. Szpektor. When web search fails, searchers become askers: Understanding the transition. In SIGIR, pp. 801–810 (2012).

[17] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: A survey study of status message q&a behavior. In CHI, pp. 1739–1748 (2010).

[18] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama, and K. Tanaka. Trustworthiness analysis of web search results. In ECDL, pp. 38–49 (2007).

[19] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In SIGIR, pp. 225–231 (2001).

[20] H. Wang, Y. Song, M.-W. Chang, X. He, A. Hassan, and R. W. White. Modeling action-level satisfaction for search task satisfaction prediction. In SIGIR, pp. 123–132 (2014).

[21] R. White. Beliefs and biases in web search. In SIGIR, pp. 3–12 (2013).

[22] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In WSDM, pp. 132–141 (2009).

[23] R. W. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems*, 27(4), pp. 23:1–23:37 (2009).

[24] R. W. White and R. A. Roth. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1), pp. 1–98 (2009).

[25] Y. Xu and D. Mease. Evaluating web search using task completion time. In SIGIR, pp. 676–677 (2009).

[26] Y. Yamamoto and K. Tanaka. Enhancing credibility judgment of web search results. In CHI, pp. 1235–1244 (2011).

[27] E. Yilmaz, M. Verma, N. Craswell, F. Radlinski, and P. Bailey. Relevance and effort: An analysis of document utility. In CIKM, pp. 91–100 (2014).